R-838-PR

November 1971

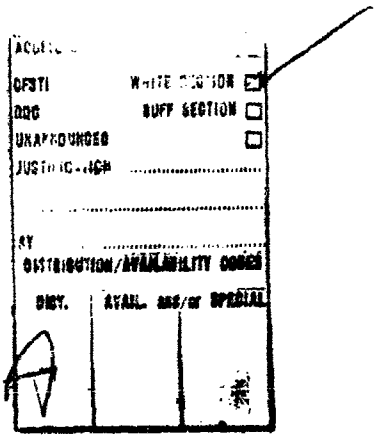
Adaptive Policies for Markov Renewal Programs

Bennett L. Fox and John E. Rolph

A Report prepared for

UNITED STATES AIR FORCE PROJECT RAND

Rand
SANTA MONICA, CA. 90406



This research is supported by the United States Air Force under Project Rand—Contract No. F44620-67-C-0045—Monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this study should not be interpreted as representing the official opinion or policy of Rand or of the United States Air Force.

DOCUMENT CONTROL DATA

1. ORIGINATING ACTIVITY The Rand Corporation	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP —

3. REPORT TITLE

ADAPTIVE POLICIES FOR MARKOV RENEWAL PROGRAMS

4. AUTHOR(S) (Last name, first name, initial)

Fox, Bennett L. and John E. Rolph

5. REPORT DATE	6a. TOTAL NO. OF PAGES	6b. NO. OF REFS.
November 1971	26	16

7. CONTRACT OR GRANT NO.	8. ORIGINATOR'S REPORT NO.
F44620-67-C-0045	R-838-PR

9a. AVAILABILITY/LIMITATION NOTICES	9b. SPONSORING AGENCY
	United States Air Force Project RAND

10. ABSTRACT

Recasts a class of infinite-state, infinite-action Markov renewal programs with unknown parameters as one-state programs with actions corresponding to stationary policies in the original program. Under suitable conditions, an adaptive (nonstationary) optimal policy is found in the sense of maximizing long-run expected reward per unit time.

11. KEY WORDS

Markov Processes
Probability
Inventory Control
Queueing
Logistics

PREFACE

Inventory, maintenance, and queuing models are the lifeblood of Air Force logistics research. Markov renewal programs are imbedded in most of those that are amenable to analytic treatment. When they are explicitly recognized, the analysis is often streamlined. Contrary to usual practice, we do not make the (generally unrealistic) assumption that all parameters of the model are known. However, information about them is acquired sequentially by observing the reward stream, transition times, and successive states visited, which depend on the policy employed.

For maximizing long-run average reward, we find a history-remembering, adaptive policy that does as well as we could if we knew all the parameters.

SUMMARY

We recast a class of infinite-state, infinite-action Markov renewal programs with unknown parameters as one-state programs with actions corresponding to stationary policies in the original program. Under suitable conditions we find an adaptive (nonstationary) optimal policy in the sense of maximizing long-run expected reward per unit time.

ACKNOWLEDGMENTS

We are indebted to Colin Mallows and Ralph Strauch, whose careful reading of earlier versions of this Report led to substantial improvements.

CONTENTS

ADAPTIVE POLICIES FOR MARKOV RENEWAL PROGRAMS

1. INTRODUCTION

Finite-state, finite-action Markov renewal programs with all parameters known were defined by Jewell in [9]. We study the infinite-state, infinite-action analog with the parameters unknown. One-step reward distributions, transition time distributions, and transition probabilities are unknown a priori. Beginning in state $i \in S$, the decision-maker takes action $k \in A_i$, moves to state j with probability p_{ij}^k, and given that he moves to j, receives reward R_{ij}^k during a transition lasting T_{ij}^k after which he takes another action, has a transition, etc. His objective is to find a policy which maximizes his expected long-run average reward. A policy $(\delta_1, \delta_2, \ldots)$ is a collection of functions mapping states into actions. At the n-th decision (transition), $\delta_n(i)$: $i \to A_i$ where δ_n may depend on the history of the process prior to n. A stationary policy δ is of the form $(\delta, \delta, \ldots)$; it uses the same function for each decision and thus cannot be history-remembering. Define Δ to be the set of all stationary policies. Making certain assumptions, we construct a nonstationary adaptive policy which does as well as any stationary policy in maximizing expected reward per unit time no matter what values the unknown parameters have.

Thus, using the average reward rate criterion, our policy is optimal whenever a stationary optimal or stationary

ϵ-optimal policy exists. Lippman [10] shows the existence of a stationary ϵ-optimal policy under essentially our assumptions, although in general a stationary optimal policy need not exist. Similar results are unattainable in the general multichain case because there is no way to be sure of optimizing the action in a transient state when the action determines which absorbing chain will be entered. A stationary optimal policy exists in the finite-state, finite-action case (Fox [5]), and sufficient conditions for existence have been given (Fox [6]) in the finite-state, infinite-action case.

Mallows and Robbins [12] give results analogous to ours in the discrete-time, one-state case. Part of our argument is an adaptation of theirs. Baños [1] and Shubert [16] treat similar problems from game theoretic and statistical decision theoretic viewpoints, respectively.

2. THE MAXIMIZING POLICY

Suppose a stationary policy $\delta \in \Delta$ is always used. For any fixed path let

$$R^{\delta}(t) = \text{total reward received up to time } t.$$

The strong law of large numbers together with a standard renewal theory argument imply that

$$\lim_{t\to\infty} \frac{1}{t} R^{\delta}(t)$$

converges to a constant with probability 1. Thus the expected gain rate associated with δ is defined as

$$g^{\delta} = \lim_{t\to\infty} \frac{1}{t} E[R^{\delta}(t)].$$

Let $g^{*} = \sup_{\delta\in\Delta} g^{\delta}$. Using the adaptive policy for any fixed path, let

$$R(t) = \text{total reward received up to time } t.$$

Under the assumptions given below we show that the rewards from the adaptive policy satisfy

$$(1) \qquad P[\lim_{t\to\infty} t^{-1} R(t) = g^{*}] = 1.$$

It will then follow that

$$(2) \qquad \lim_{t\to\infty} t^{-1} E[R(t)] = g^*.$$

The assumptions:

1. There is an a priori known countable set of stationary policies $\Lambda \subset \Delta$ such that

$$\sup\{g^\lambda: \lambda \in \Lambda\} = g^*.$$

2. For each state, the expectation and the variance of the time and reward until state 1 is next reached is uniformly bounded over Λ.

3. For each state, the expected time to return to state 1 is uniformly bounded away from 0 over Λ.

Note that uniform bounds over $S \times \Lambda$ are not needed. Our proof uses assumptions 1, 2, and 3 directly. In [15] it is shown that assumptions 2 and 3 on Δ rather than Λ imply assumption 1, thus effectively eliminating the need for assumption 1.

Alternatively, conditions on the transition matrices, one-step reward distributions, and one-step time distributions can be given which imply assumptions 2 and 3. They are:

2'. The means and variances of the one-step times and rewards are uniformly bounded from above over actions and states.

3'. The mean of the one-step times is uniformly bounded away from zero over states and actions.

4'. The semi-Markov process associated with any stationary policy is regular and the mean and variance of the time to get to state 1 is uniformly bounded over $S \times \Lambda$.

To ensure that our adaptive policy will concentrate with probability 1 on the high gain rate stationary policies, only policies in Λ are tried and each policy in Λ is tried infinitely often. The basic unit used in defining our policy is the state 1-to-state 1 cycle. Within each state 1-to-state 1 cycle the same stationary policy is used.

Beginning in the initial state some fixed stationary policy is applied until state 1 is reached. From this point forward we have a one-state problem since policy choices are only made at state 1. Following Mallows and Robbins [12] our strategy specifies a sequence of positive integers which number the forced-choice cycles in which predetermined stationary policies are applied. Let $s_{11}, s_{12}, \ldots$ be any increasing sequence of positive integers with $s_{11} = 1$. Let $s_{21}, s_{22}, \ldots$ be a second disjoint sequence with $s_{31}, s_{32}, \ldots$ being a third sequence disjoint from the first two, and so on. If for the n-th cycle $n = s_{\delta t}$, for some δ, t we use stationary policy δ for the n-th cycle; otherwise we choose the stationary policy with the leading

observed reward rate. The observed reward rate for policy δ at time t is given by

$$R^{\delta}(t) = \frac{R^{\delta}(t)}{B^{\delta}(t)} \quad ,$$

where $R^{\delta}(t)$ is the total reward received and $B^{\delta}(t)$ is the total time spent prior to t while δ was applied. The relationship is defined only for those δ which have been applied in a forced-choice cycle prior to t. Let s(n) be the total number of forced choices up to the n-th cycle, i.e., the number of integers $s_{\delta\ell}$ which are $\leq$ n. We choose the $s_{\delta\ell}$ so that

$$\sum_{n=1}^{\infty} \left(\frac{s(n)}{n}\right)^2 < \infty.$$

A choice which satisfies this is $s(n) \doteq \log n$.

In sampling policies directly rather than actions, we do not fully utilize all information since each action is associated with many different policies. We do not know a general remedy, but in Section 4 we give a modified policy for the finite case that uses information more efficiently.

3. PROOF OF OPTIMALITY

Let the random variables U^δ and V^δ be, respectively, reward and time for a state 1-to-state 1 cycle using δ. In view of assumption 2, U^δ and V^δ have expectations and variances uniformly bounded over Λ. Call this bound H. The policy we use essentially reduces the original problem to a one-state problem, where the transition times need not be constants. As in [12], we can define a nondecreasing sequence $\{c_n\}$ such that

$$(3)\quad c_n > 0,\quad \sum_n c_n^{-2} < \infty,\quad n^{-1}s(n)\,c_n \to 0.$$

A particular choice which can be shown to satisfy (3) is $c(n) = \frac{n}{s(n)} R_n^{\frac{1}{2}}$ where $R_n = \sum_{j=n}^{\infty} \left(\frac{s(j)}{j}\right)^2$. Denote the distribution of U^δ by F^δ. By the Markov inequality

$$F^\delta(-c_n) \le H\, c_n^{-2}, \quad \delta \in \Lambda .$$

Hence, if D_n is the policy used on the n-th cycle and U_n is the corresponding reward, then

$$P\{U_n \le -c_n\} = \sum_{\delta \in \Lambda} P(D_n = \delta) F^\delta(-c_n) \le Hc_n^{-2} ,$$

and so $\sum_n P\{U_n \le -c_n\}$ converges by (3). By the Borel-Cantelli lemma, $U_n \le -c_n$ only finitely often w.p.1; so w.p.1 there is an N_1 such that $U_n > -c_n$ for all $n > N_1$. A similar

argument shows that there is an N_2 such that w.p.1 $|U_n| < c_n$ and $V_n < c_n$. Hence, by (3), we can neglect the contribution of the forced-choice cycles to the overall reward and time.

At time t, define $f_1(t)$ as the number of forced choices prior to time t and $f_2(t)$ as the number of different policies used on free choices prior to t. Let p(t) = the number of different policies used prior to time t. Then

$$f_1(t) \geq p(t) \geq f_2(t).$$

Define the last free-choice cycle prior to t for a policy δ as the last cycle, if any, for which policy δ was chosen as the leader. By the above argument the contribution of the last free-choice cycles can be neglected. Assumption 2 implies that the time and rewards before reaching state 1 for the first time can also be neglected.

Indexing consecutive cycles by m and excluding the time and rewards before reaching state 1 for the first time we define

V_i = time to complete i-th cycle,

$$V_i^\delta = \begin{cases} \text{time to complete i-th cycle if policy } \delta \text{ is used,} \\ 0 \text{ otherwise.} \end{cases}$$

$$B(m) = \sum_{i=1}^{m} V_i$$

$$B^\delta(m) = \sum_{i=1}^{m} V_i^\delta \quad .$$

U_i, U_i^δ, $R(m)$, $R^\delta(m)$ are defined similarly for the reward sequence. Let

$N^\delta(m)$ = the number of the first m cycles which use policy δ.

Fix a policy $\sigma \in \Lambda$;

$$\bar{R}^\sigma(m) = \frac{R^\sigma(m)}{B^\sigma(m)} = \frac{R^\sigma(m)/N^\sigma(m)}{B^\sigma(m)/N^\sigma(m)} .$$

By the SLLN the numerator and denominator converge to a constant w.p.1 as $N^\sigma(m) \to \infty$. Since the union of two null sets is null and $N^\sigma(m) \to \infty$ as $m \to \infty$, the advanced calculus result that the limit of a ratio is the ratio of the limits yields

$$(4) \quad \lim_{m\to\infty} \bar{R}^\sigma(m) = g^\sigma \quad \text{w.p.1.}$$

This is a special case of a result of Pyke and Schaufele [14, Theorem 5.1]. Strictly speaking, the fact that the constant is g^σ depends on the lemma proved later that relates cycles and continuous time. Using the fact that policy σ is applied infinitely often ($s_{\sigma 1}$, $s_{\sigma 2}$, ...) in forced-choice cycles, we can choose m large enough to guarantee in advance that $N^\sigma(m) \geq N$ for any fixed N. Thus fixing ϵ_1, $\epsilon_2 > 0$, choose m_o such that

$$P\{\bar{R}^{\sigma}(m) \geq g^{\sigma} - \epsilon_2 \text{ for all } m > m_o\} \geq 1 - \epsilon_1,$$

from the above SLLN argument.

Let Γ be the set of policies used in free choices for $m \geq m_o$. For $\gamma \in \Gamma$ and $m \geq m_o$, define $\iota^{\gamma}(m)$ as the largest cycle index $\leq m$ where γ was freely chosen. From the definitions,

$$\bar{R}^{\delta}(\iota^{\delta}(m) - 1) \geq \bar{R}^{\sigma}(\iota^{\delta}(m) - 1),$$

for all $\sigma \in \Lambda$.

By earlier arguments and neglecting the contributions from policies not used after cycle m_o,

$$\bar{R}(m) = \sum_{\gamma \in \Gamma} R^{\gamma}(\iota^{\gamma}(m) - 1)/B(m) + o_p(1),$$

where the last term goes to 0 w.p.1 as $m \to \infty$. Thus,

$$\bar{R}(m) \geq \sum_{\gamma \in \Gamma} \bar{R}^{\sigma}(\iota^{\gamma}(m) - 1)B^{\gamma}(\iota^{\gamma}(m) - 1)/B(m) + o_p(1)$$

$$\geq (g^{\sigma} - \epsilon_2)\left[\sum_{\gamma \in \Gamma} B^{\gamma}(\iota^{\gamma}(m) - 1)/B(m)\right] + o_p(1)$$

$$\geq g^{\sigma} - \epsilon_2 + o_p(1),$$

since the term in square brackets goes to 1 w.p.1 as $m \to \infty$, and as ϵ_2 was arbitrary,

$$\lim_{m\to\infty} \inf \bar{R}(m) \geq g^{\sigma} \quad \text{w.p.1.}$$

Since σ was arbitrary and a denumerable union of null sets is null,

$$(5) \qquad P(\lim_{m\to\infty} \inf \bar{R}(m) \geq g^{*}) = 1.$$

Now $\bar{R}^{\delta}(m) \to g^{\delta}$ w.p.1 as $m \to \infty$ so that

$$(6) \qquad P(\lim_{m\to\infty} \sup \bar{R}(m) \leq g^{*}) = 1,$$

since $\bar{R}(m)$ is a weighted average of the $\bar{R}^{\delta}(m)$. Thus we have proved

$$(7) \qquad P(\lim_{m\to\infty} \bar{R}(m) = g^{*}) = 1.$$

To complete the proof of (1) we need a lemma to allow us to move from indexing by number of transitions to indexing by time.

LEMMA.

(a) $\lim_{m\to\infty} \sup \frac{B(m)}{m} < \infty$ w.p.1.

(b) $\lim_{m\to\infty} \inf \frac{B(m)}{m} > 0$ w.p.1.

Proof. For the proof of the lemma a more abstract representation of the process is required. The process can be viewed as a sequence of functions on an underlying measure space. See [11] or [15].

Define the σ-field

$$\mathfrak{F}_j = \sigma\{D_i, V_i, U_i; i = 1, 2, \ldots, j\}$$

where $\sigma\{\cdot\}$ means the smallest σ-field over which $\{\cdot\}$ is measurable. Thus $E(V_i|\mathfrak{F}_{i-1})$ is the expected time to complete the i-th cycle given the previous history of the process while $E(V_i|D_i)$ is the expected time to complete the i-th cycle given the value of D_i, the stationary policy used. Note that $\mathfrak{F}_j \supset \mathfrak{F}_{j-1}$, D_j is measurable over $\mathfrak{F}_{j-1}$, and

$$E(V_j|\mathfrak{F}_{j-1}) = E(V_j|D_j).$$

Clearly,

$$B(m) = \sum_{i=1}^{m} [V_i - E(V_i|D_i)] + \sum_{i=1}^{m} E(V_i|D_i).$$

Thus, since $E(V_i|D_i)$ is bounded away from 0 and ∞ by assumptions 2 and 3, it suffices to show that the first term is negligible,

$$E[(V_i - E(V_i|D_i))|D_i] = 0$$

and

$$\sum_{i=1}^{\infty} E[(V_i - E(V_i|D_i))^2]/i^2$$

$$\leq H \sum_{i=1}^{\infty} 1/i^2 < \infty .$$

By a standard martingale theorem (Feller [4, pp. 234-238]),

$$\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} (V_i - E(V_i|D_i)) = 0 \quad \text{w.p.1}\|.$$

Turning to proof of (2), we let P be a measure on the reward sequence corresponding to our policy. For any $a > 0$,

$$\int_{|\bar{R}(t)|>a} |\bar{R}(t)|dP \leq a \int_{|\bar{R}(t)|\geq a} [\bar{R}(t)/a]^2 dP \leq H/a \to 0 \text{ as } a \to \infty.$$

Thus, the random variables $\{\bar{R}(t)\}$ are uniformly integrable so (1) implies (2). (Loève [11], p. 163.) We have proved the following.

THEOREM. Under assumptions 1, 2, and 3, the average rewards from the above strategy satisfy

(i) $P\{\lim_{t\to\infty} t^{-1}R(t) = g^*\} = 1$

(ii) $\lim_{t\to\infty} t^{-1} E[R(t)] = g^*$.

4. REMARKS

In the finite-state, finite-action case we could sample actions on each transition rather than policies on each cycle. By sampling actions on forced-choice transitions we can obtain consistent estimates of the parameters. One choice is the natural empirical estimators which are consistent (Moore and Pyke [13]). On the free-choice transitions, we follow the leader obtained by substituting these estimates for the unknown parameters in the gain rate formula for stationary policies. The estimated optimal policy can be calculated using a policy improvement routine or linear program (Fox [5], Denardo and Fox [2]). The proof that g^* is attained is different than the above one since we do not reduce the problem to one state, but the number of policies being finite leads to some simplifications. In the finite case, the existence of a stationary optimal policy makes our policy optimal. Intuition indicates a faster convergence rate for sampling actions directly rather than just sampling policies.

Many problems studied in the literature satisfy our assumptions; for example,

(i) replacement problems where we return to state 1 (replace the item) whenever the state (or deterioration) exceeds a certain level, to be determined.

(ii) queuing problems where we activate the server whenever the queue length exceeds a certain level, to be determined. Heyman [8] gives conditions under which a policy of this form is optimal for the M/G/1 queue. To satisfy our assumptions, we rule out policies that do not activate the server when the queue length exceeds a given (large) number.

(iii) inventory problems were we determine a reorder point and a reorder level. See Hadley and Whitin [7, Chap. 8].

Lippman [10] mentions another example: the "streetwalker's dilemma," where the server must decide whether to accept a given proposition or wait for a more desirable one. He gives simple conditions under which our assumptions hold and the optimal policy has the form: accept an offer if and only if the ratio of expected reward to expected service time exceeds a certain number.

REFERENCES

1. Baños, A., "On Pseudo-Games," Ann. Math. Stat., 39, 1968, pp. 1932-1945.

2. Denardo, E. V., and B. L. Fox, "Multichain Markov Renewal Programs," SIAM J. Appl. Math., 16, 1968, pp. 468-487.

3. Doob, J. L., Stochastic Processes, John Wiley and Sons, New York, 1952.

4. Feller, W., An Introduction to the Theory of Probability and Its Applications, Vol. 2, John Wiley and Sons, New York, 1966.

5. Fox, B. L., "Markov Renewal Programming by Linear Fractional Programming," SIAM J. Appl. Math., 14, 1966, pp. 1418-1430.

6. Fox. B. L., "Existence of Stationary Optimal Policies for Some Markov Renewal Programs," SIAM Review, 9, 1967, pp. 573-576.

7. Hadley, G., and T. M. Whitin, Analysis of Inventory Systems, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

8. Heyman, D. P., "Optimal Operating Policies for M/G/1 Queuing Systems," Operations Res., 16, 1968, pp. 362-382.

9. Jewell, W. S., "Markov Renewal Programming, I and II," Operations Res., 11, 1963, pp. 938-971.

10. Lippman, S. A., Maximal Average Reward Policies for a Class of Semi-Markov Decision Processes with Arbitrary State and Action Space, Western Management Science Institute Paper 162, University of California, Los Angeles, October 1970. (To appear in Ann. Math. Stat.)

11. Loève, M., Probability Theory, 3d ed., Van Nostrand Co., Princeton, New Jersey, 1963.

12. Mallows, C. L., and H. Robbins, "Some Problems of Optimal Sampling Strategy," J. Math. Anal. Appl., 8, 1964, pp. 90-103.

13. Moore, E. H., and R. Pyke, "Estimation of the Transition Distributions of a Markov Renewal Process," Ann. Inst. Stat. Math., 20, 1968, pp. 411-424.

14. Pyke, R., and R. Schaufele, "Limit Theorems for Markov Renewal Processes," Ann. Math. Stat., 37, 1964, pp. 1746-1764.